US Airbnb Listing Price Recommender

Sophie Lou, Sophia Jiang, Ziyue Li, Jianing Yu

Dec 15th, 2021

Our Team

Sophie Lou

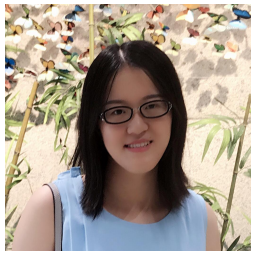
Senior,
Data Science

Sophia Jiang

Senior,
Data Science & Econ

Ziyue Li

Senior,
Data Science & Econ



Jianing Yu

Senior,
Data Science & Econ

Table of Contents

1. Introduction

Negative effects of Covid-19 on Airbnb

Decreasing Price

In 2020, Airbnb hosts have dropped their daily rates as much as $90 on average.

High Cancellation rate

64% of guests either have cancelled or plan to cancel an Airbnb booking since the pandemic started

Dim Future

45% of hosts won't be able to sustain operating costs if the pandemic lasts another 6 months

They Need A New Pricing Model

Source

Motivation:
Benefits to Hosts and Customers

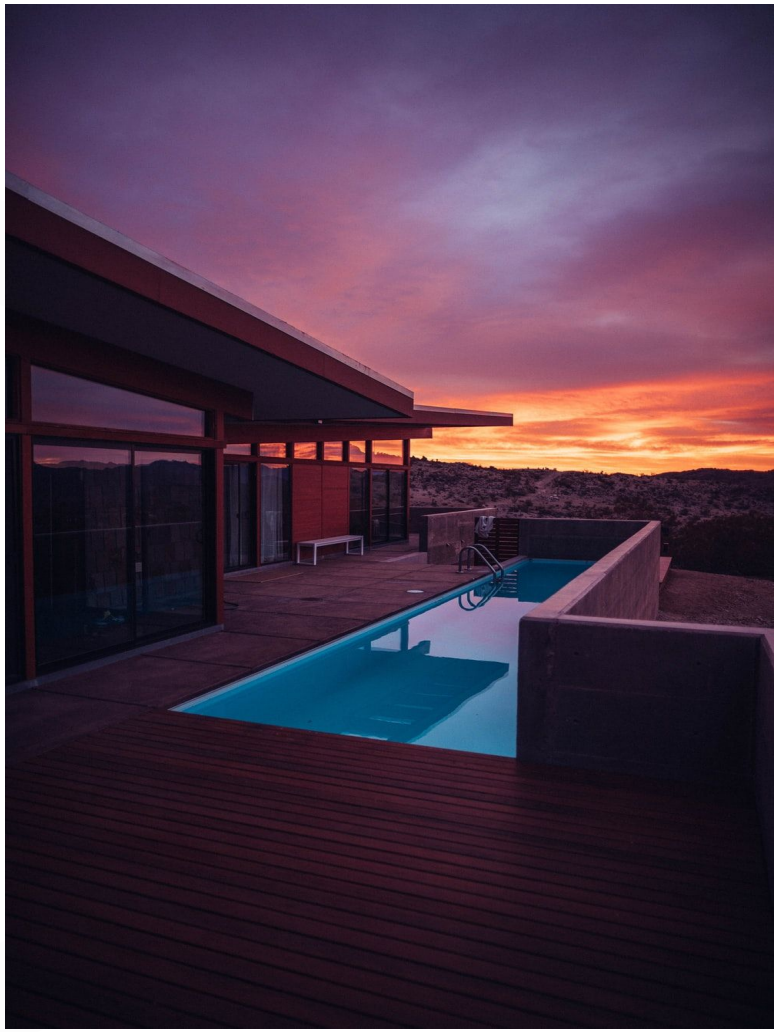
Consumers

The study allows consumers to know the market price of current airbnb housing, given the demand, location and property types. This gives customers a rational estimate of their accomodation costs.

Hosts

The suggested pricing gives the hosts a fair view of the market price. This benefits them from setting optimal price of their houses to keep their business running during the pandemic.

Research Question



How do we recommend pricing for each airbnb listing based on their location, property type, reviews, amenities, cancelation policy, etc?

Which factors are important in determining the pricing? Does it vary by state and market?

What is the real-world implementation of our pricing model?

2. Dataset Description

Airbnb Dataset

- The dataset is downloaded from opendatasoft worldwide airbnb-listings dataset. For this project, we only included US airbnb listings that are available 365 days per year, which results in 7,961 data points
- The original dataset has 96 columns. After inspecting the dataset, we dropped irrelevant columns
- Dataset structure
  - every row is a listing from Airbnb
  - 19 features
    - 11 numeric features
    - 8 categorical features

Source

Numeric Features

host_since

date the host/user was created

host_total_listings_count

number of listings the host owns

accommodates

maximum capacity of the listing

bathrooms, bedrooms, beds

number of bathrooms, bedrooms and beds in the listing

minimum_nights, maximum_nights

minimum/maximum number of night available to stay for the listing

number_of_reviews, review_scores_rating, reviews_per_month

number of reviews the listing receives, average rating for the listing, number of reviews the listing receives per month

Categorical Features

neighbourhood_cleansed: the neighborhood that the airbnb listing is located at

city: the city that the airbnb listing is located at

state: the state that the airbnb listing is located at

market: the greater area that the airbnb listing is located at

property_type: self selected property type (Hotels, Apartment, Bed and Breakfasts, etc.)

room_type: Shared room, Private room, Entire home/apt, etc.

amenities: TV, Wireless Internet, Air conditioning, Free parking, etc

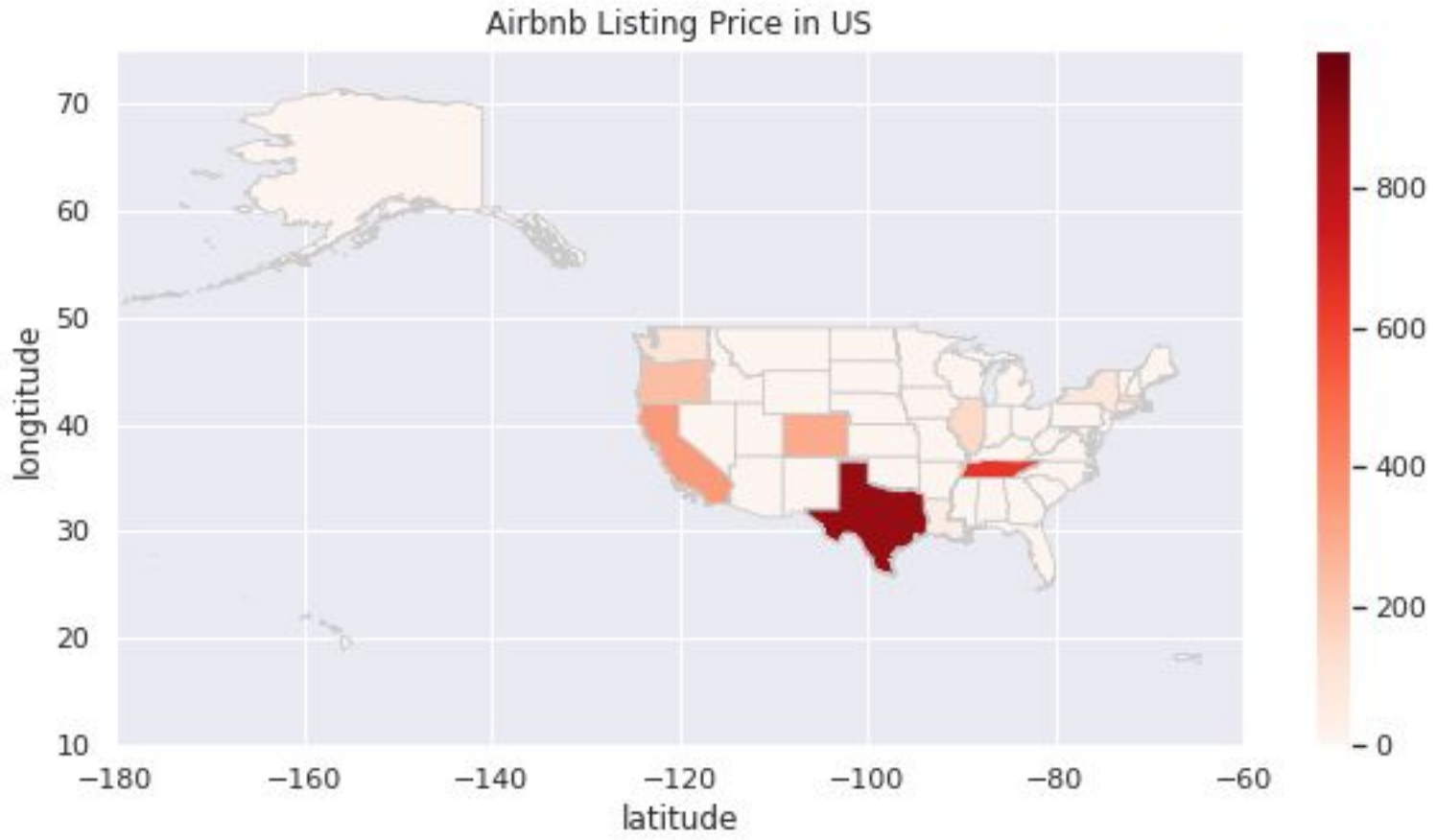
cancellation_policy: Flexible, moderate, strict

3. Data Exploration

Data Cleaning & Feature Engineering

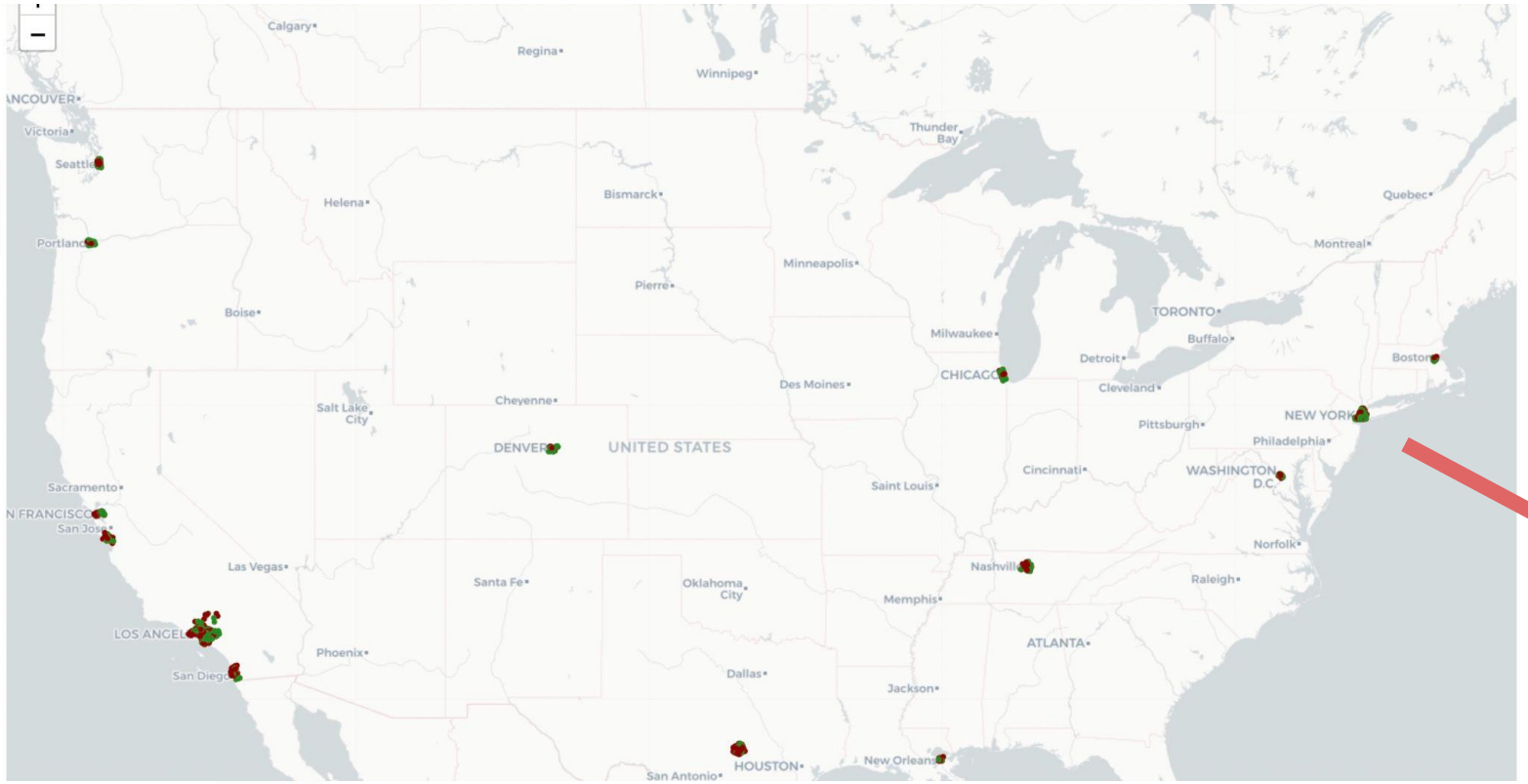
- For **Numeric variables**,
  - imputed missing value with mean based on categories (e.g. mean based on host_id, mean based on city, and etc) if possible
  - for rows that still have missing values after imputation, we dropped them
- For **Categorical variables**,
  - used one-hot encoding to convert them into new columns with binary values
- We also feature engineered ***host_since***, and ***amenities*** column
  - For ***host_since*** (e.g. 2014-02-28),
    - extracted year and subtract it from 2022
    - created a new feature ***host_year*** with the calculated value
  - For ***amenities*** (e.g. [TV, Internet, Wireless Internet, Air conditioning, …]),
    - broke the list of unique amenities into binary columns
      - 1 indicates the listing has this amenity, and 0 indicates the listing doesn't
    - dropped amenities equipped in less than 500 airbnb listings

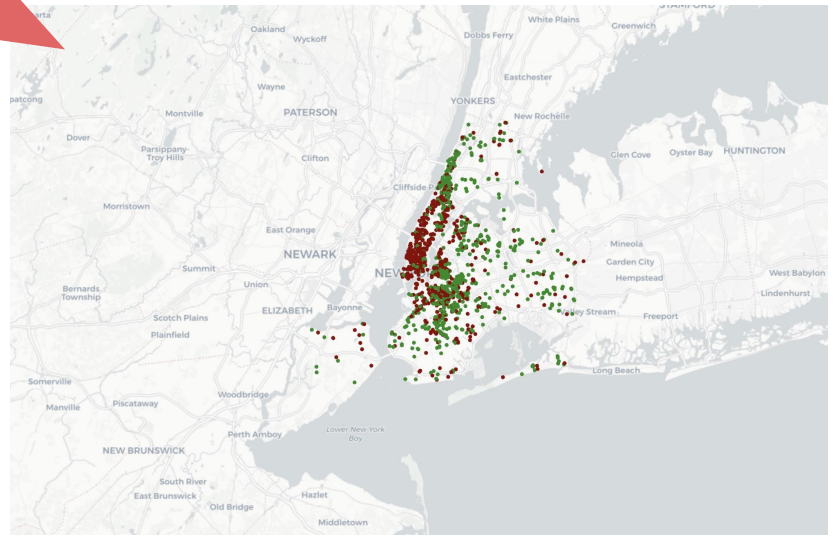
Visualization of state-wise average listing price



- ➔ We don’t have price data for many states (state with light red coloring)
- ➔ Although the states present in the dataset show differences in price, ***state*** might not be a good feature based on the model

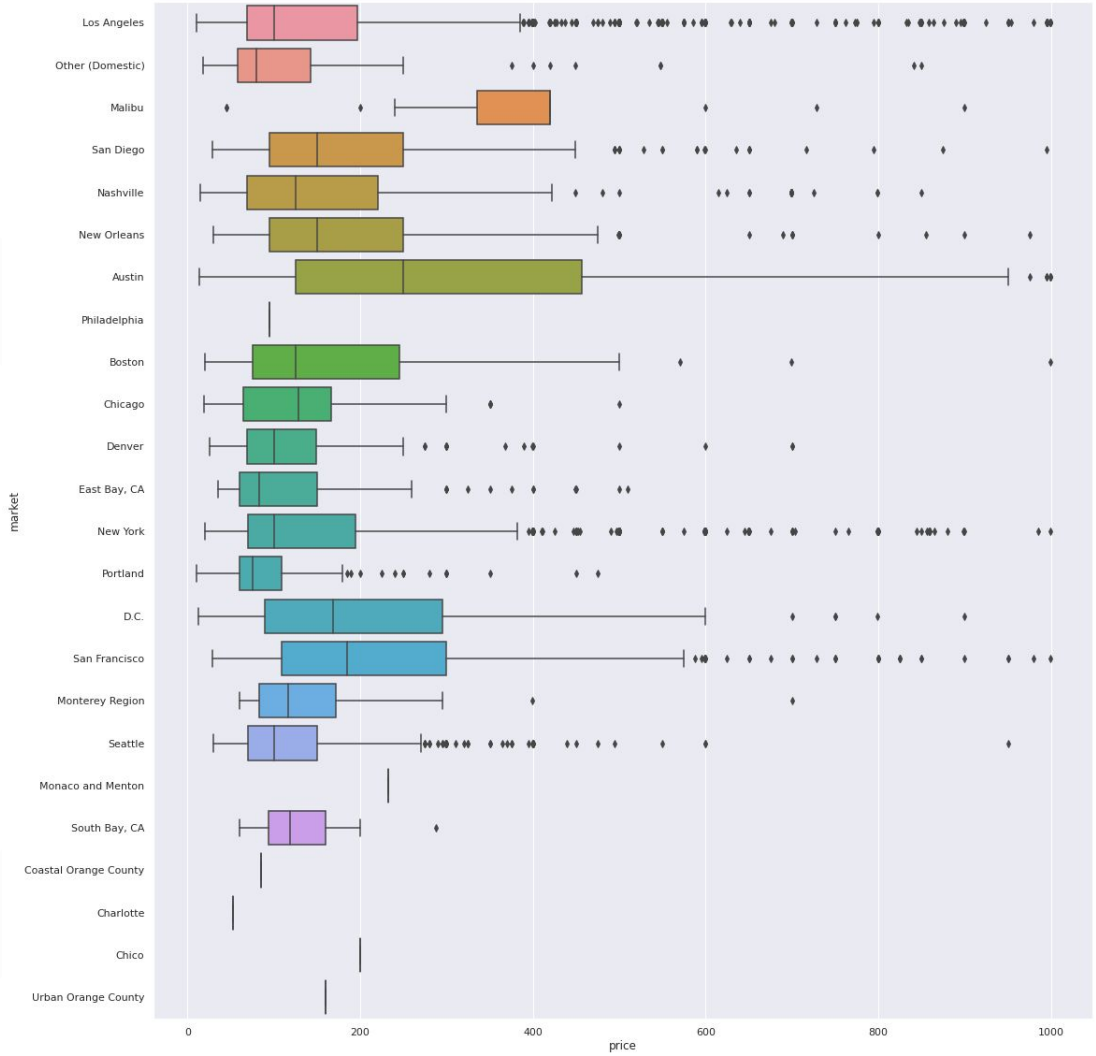
A detailed look of the listing map





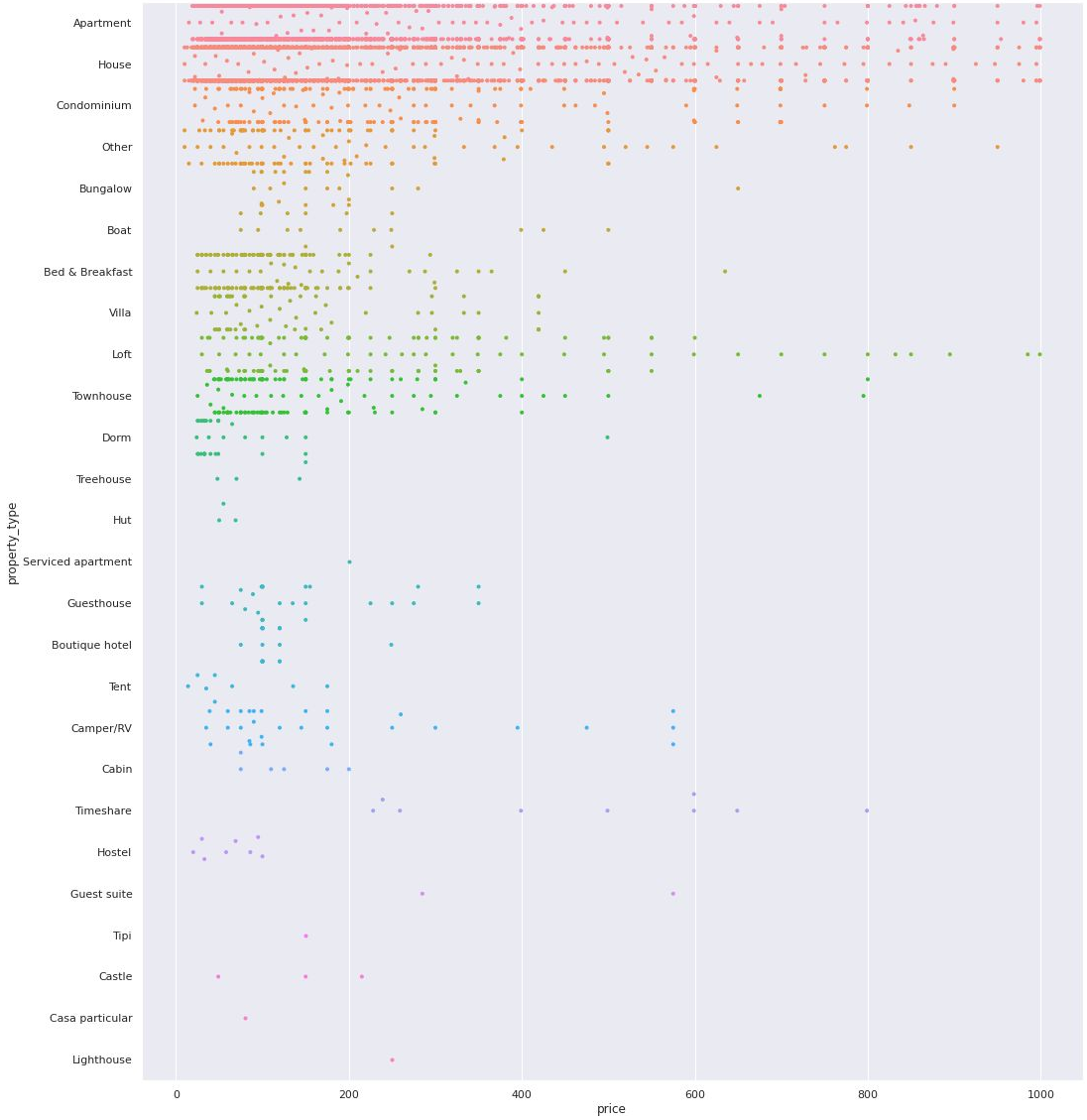
The available house listings from the dataset are clustered in greater areas including New York, Boston, Los Angeles, and etc. Hence, using one-hot on more granular category such as ***City*** and ***neighborhood-cleansed*** might not be optimal.

Visualization of market



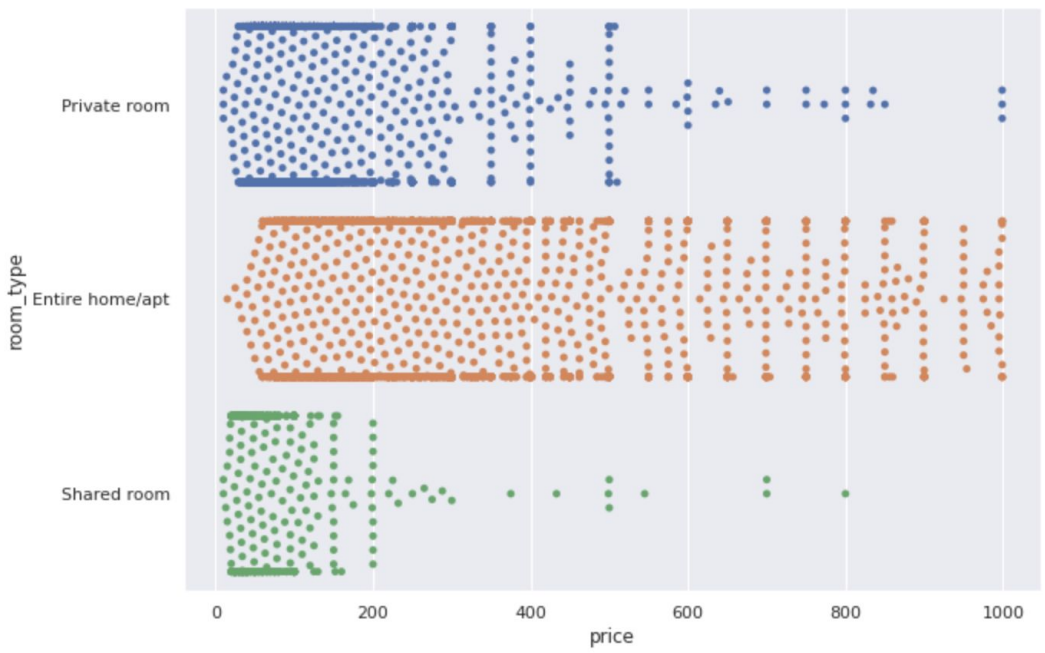
- ➔ We decided to drop ***state***, ***city***, ***neighborhood_cleansed*** and use ***Market***
- ➔ ***Market*** show difference in the distribution of price

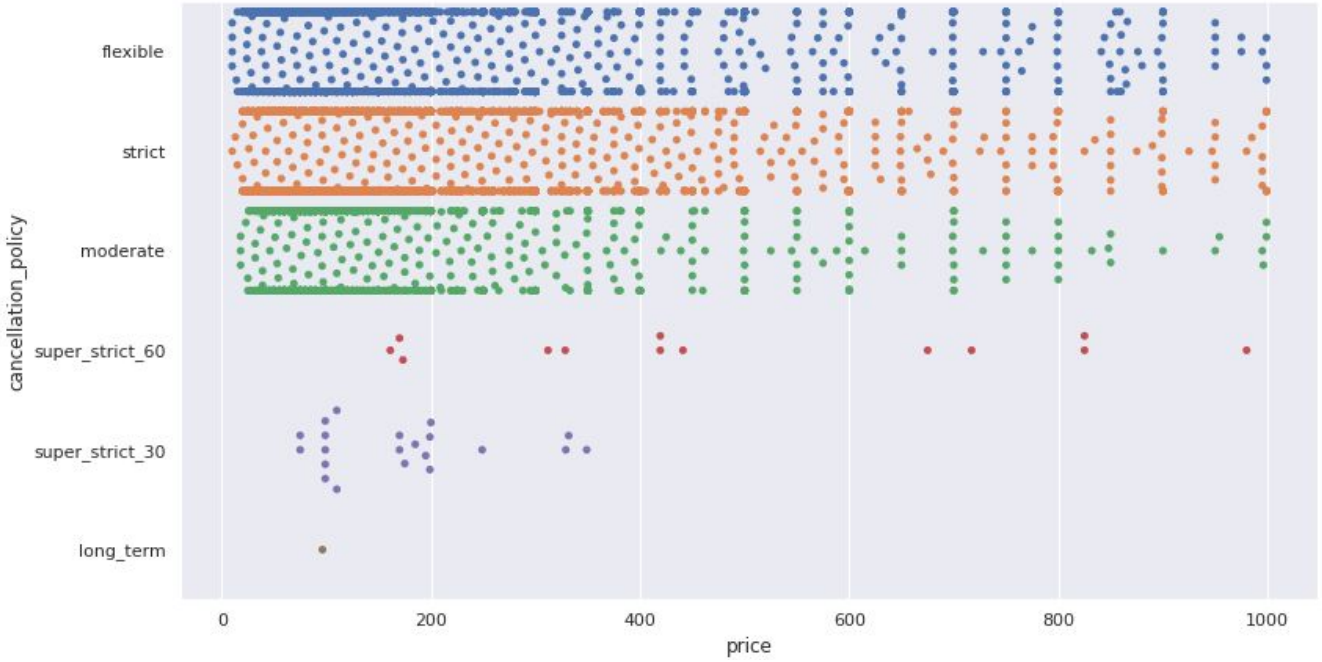
Visualization of property_type



➔ ***property_type*** have drastically different distributions of price, so we kept the ***property_type*** one hot columns

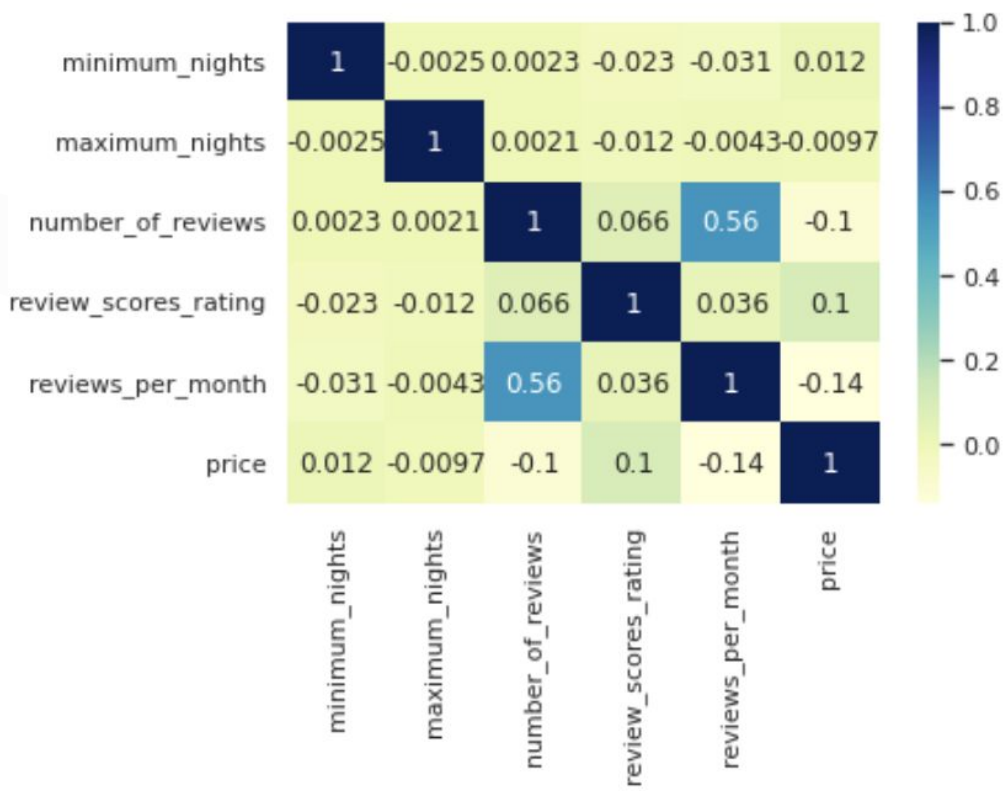
Visualization of room_type & cancellation_policy

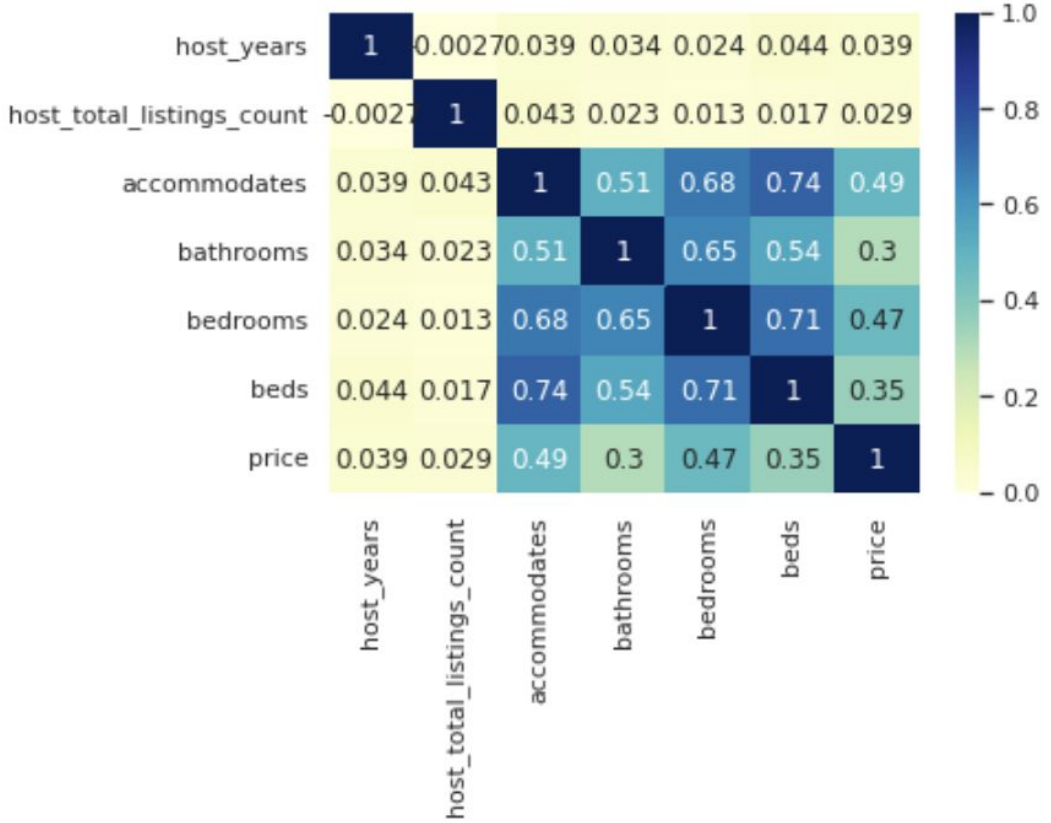




room_type *and* ***cancellation_policy*** also show variations in the distributions of price, so we kept the their one hot columns as well.

Visualization of numerical variables





price vs ***minimum_nights***: correlation is **0.012**

price vs ***maximum_nights***: correlation is **-0.0097**

drop columns with weak correlation

price vs ***host_years*:** correlation is **0.039**

price vs ***host_total_listings_count***: correlation is **0.029**

Updated Numeric Features

~~**host_since**~~

~~date the host/user was created~~

~~**host_total_listings_count**~~

~~number of listings the host owns~~

accommodates

maximum capacity of the listing

bathrooms, bedrooms, beds

number of bathrooms, bedrooms and beds in the listing

~~**minimum_nights, maximum_nights**~~

~~minimum/maximum number of night available to stay for the listing~~

number_of_reviews, review_scores_rating, reviews_per_month

number of reviews the listing receives, average rating for the listing, number of reviews the listing receives per month

Updated Categorical Features

~~**neighbourhood_cleansed:** the neighborhood that the airbnb listing is located at~~

~~**city:** the city that the airbnb listing is located at~~

~~**state:** the state that the airbnb listing is located at~~

market: the greater area that the airbnb listing is located at

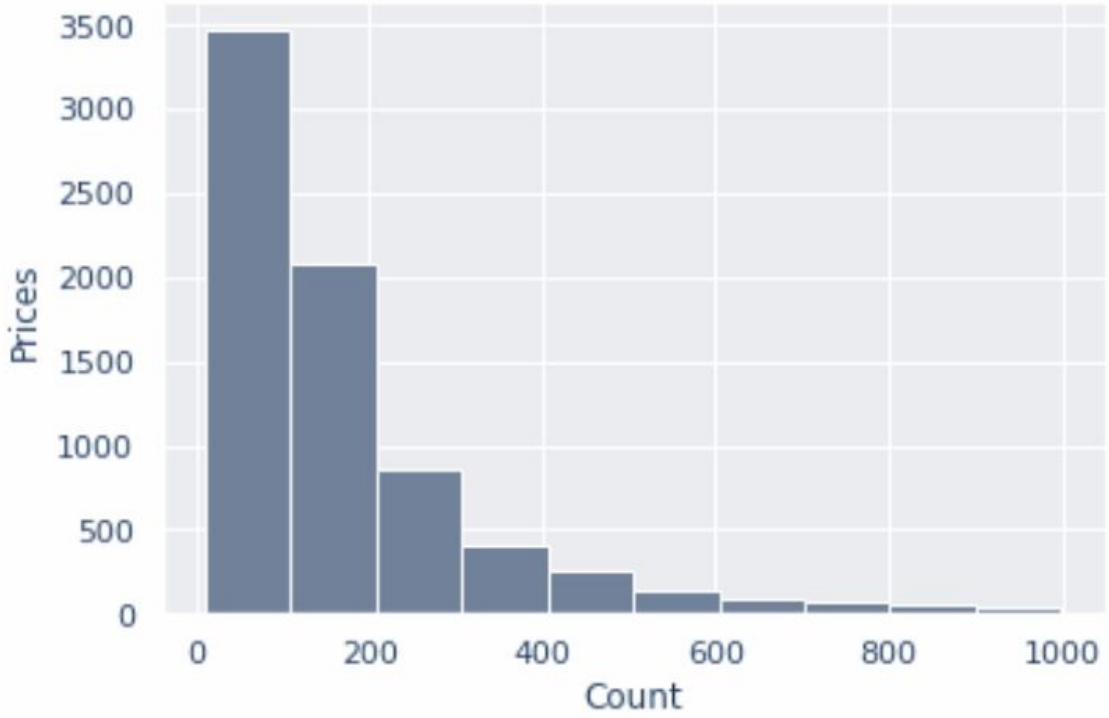
property_type: self selected property type (Hotels, Apartment, Bed and Breakfasts, etc.)

room_type: Shared room, Private room, Entire home/apt, etc.

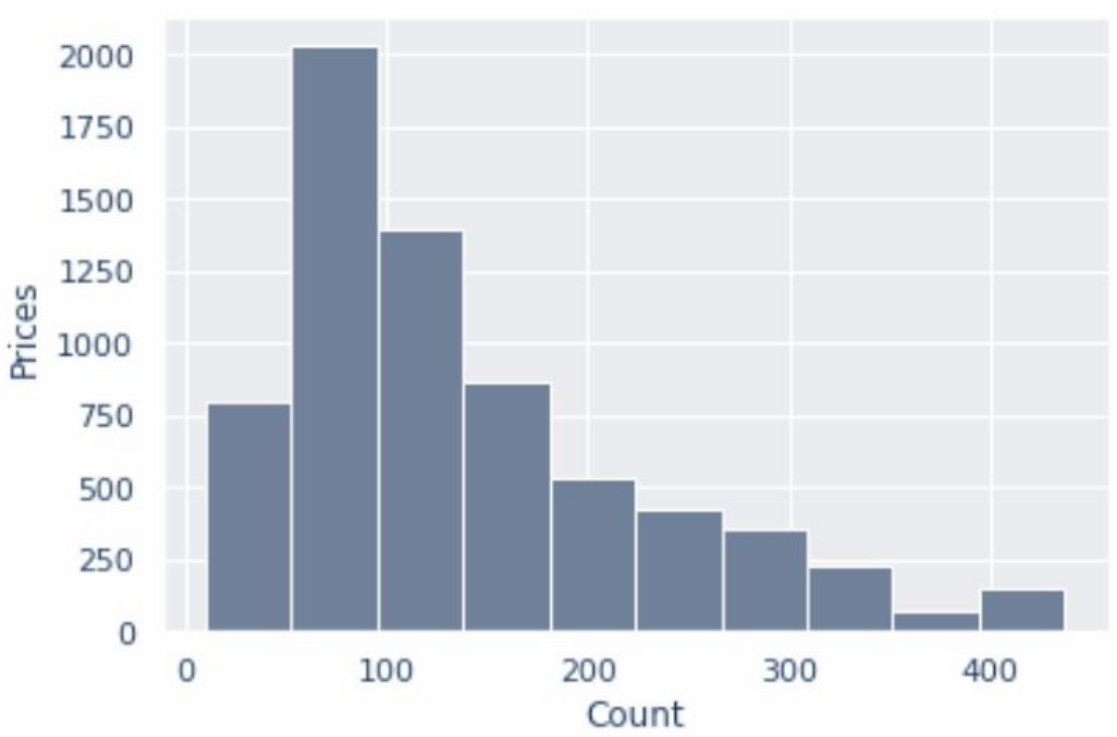
amenities: TV, Wireless Internet, Air conditioning, Free parking, etc

cancellation_policy: Flexible, moderate, strict.

Drop Outliers outside 1.5 IQR



Original Dataset



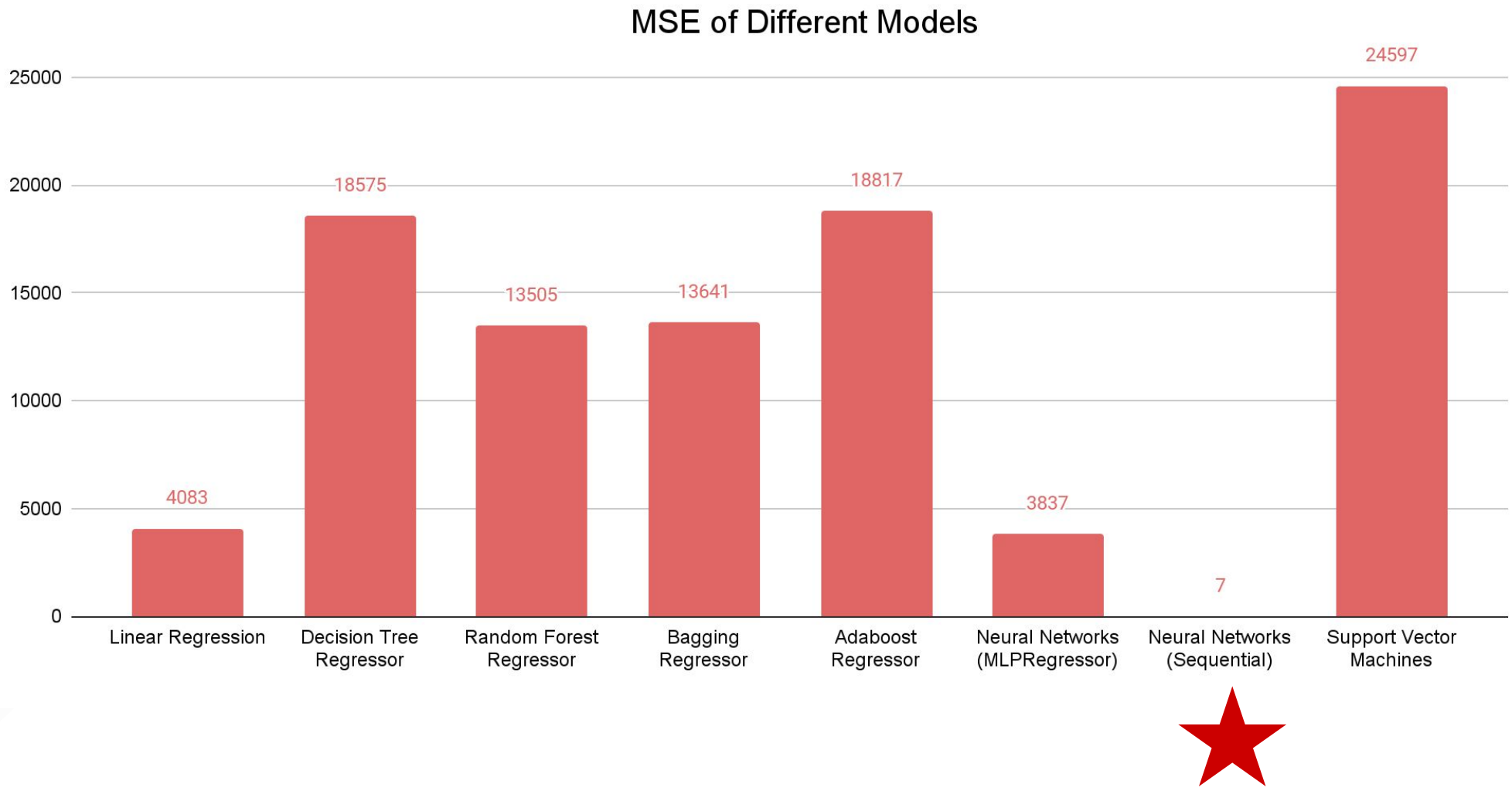
Dataset after dropping outliers

4. Machine Learning Methods

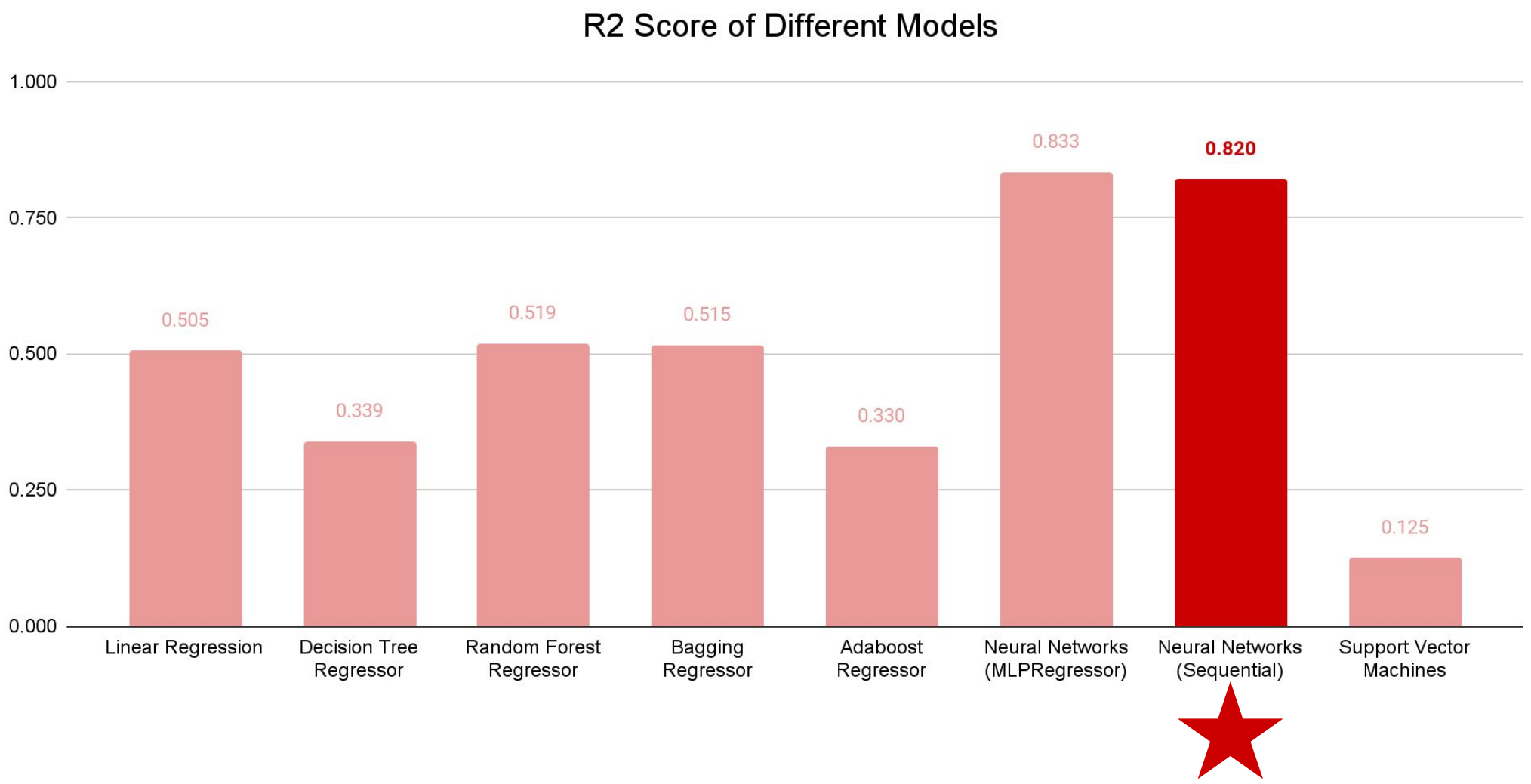
Modelling

- **Train Test Split**
  - 70% Training set, 30% Test set
- **Implement various regression models to find the best model**
  - Linear Regression
  - Decision Tree Regressor
  - Random Forest Regressor
  - Bagging Regressor
  - Adaboost Regressor
  - Regression-based Neural Networks
  - Support Vector Machines
- **Attempted optimization methods**
  - normalize data → increase MSE, lower r2 score
  - RandomizedSearch → lower MSE, increase r2 score, but take a long time to run

Model Performances Comparison



Model Performances Comparison



Sub-optimal Model: Neural Networks - MLPRegressor

We used Sklearn to implement the MLPRegressor neural network:

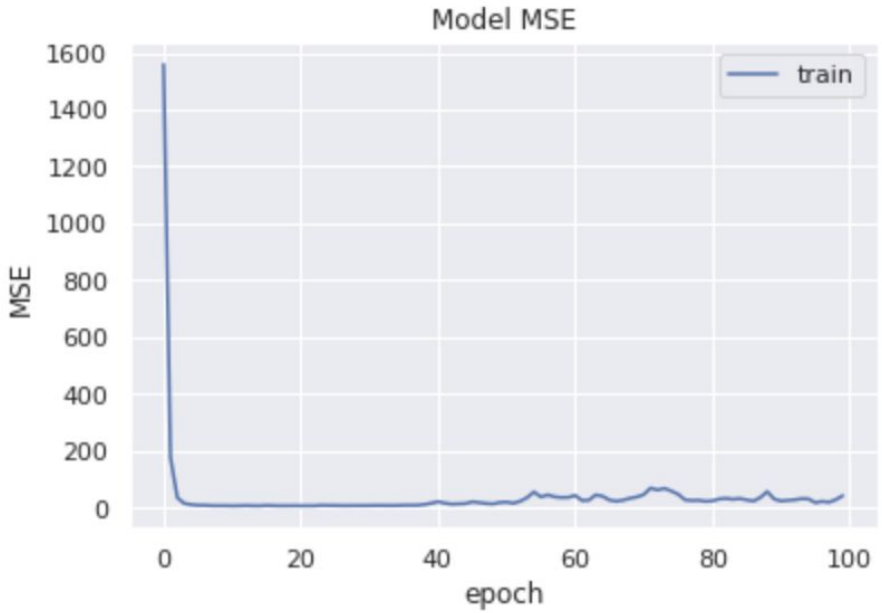
- Use (500,200) as *hidden_layer_size*
- Use **ReLu activation**
- Use *max_inter* = 50 and *random_state* = 42
- Use *learning_rate_init* = 0.1, *solver* = ‘adam’

Error:
R^2 = 0.833
MSE = 3,837

Optimal Model: Neural Networks - Sequential

We used Keras to implement the Sequential neural network:

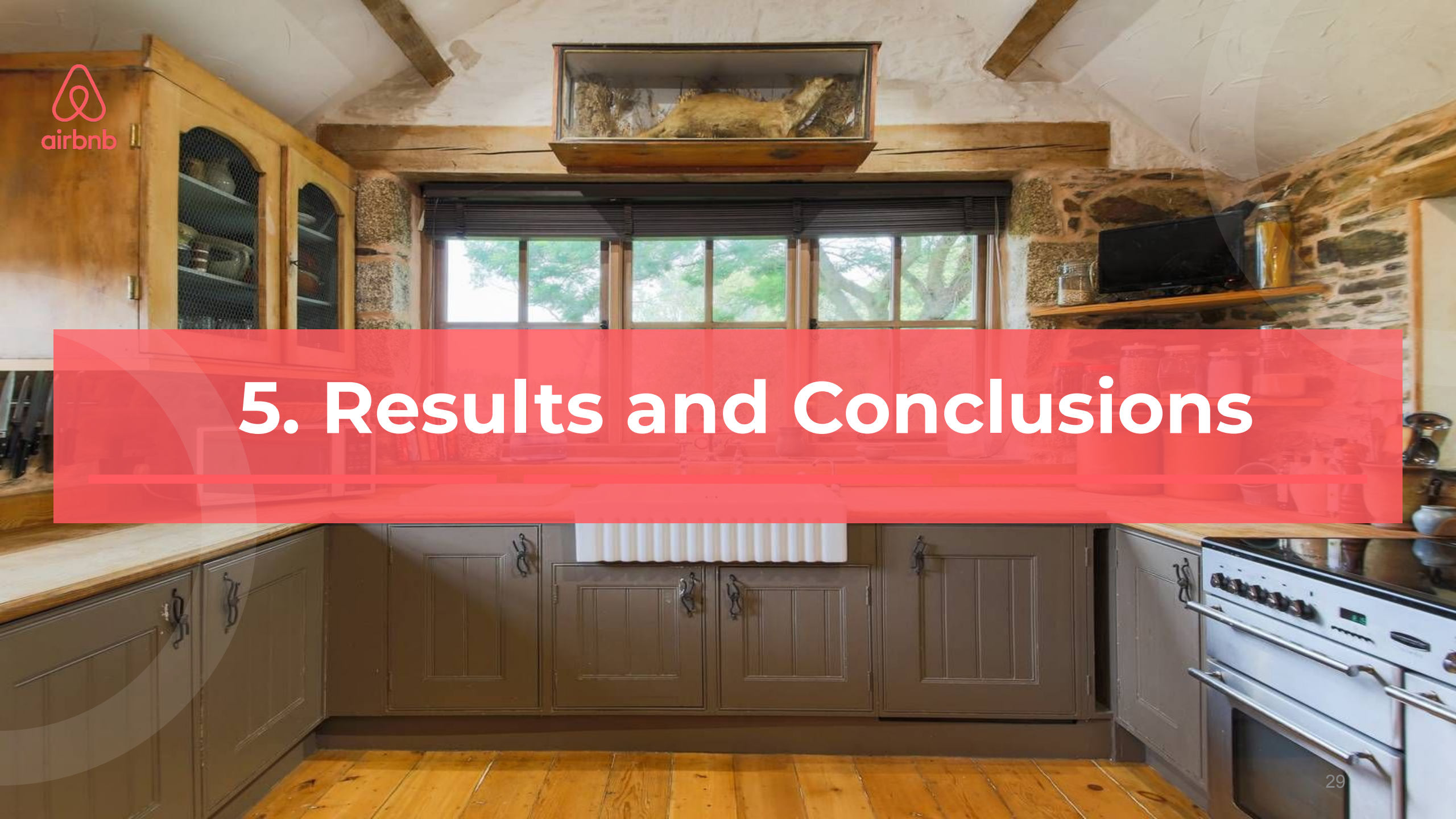
- Add *dense* = 123 and *input_dim* = 109 using **Normal Initializer** + **ReLu activation**
- Add *dense* = 2670 dense using **ReLu activation**
- Add *dense* = 1 using **Linear Activation**
- Try different **epoch numbers**, and the best possible model is with *epoch* = 100, *batch_size* = 150



Error:

R^2 = **0.820**

MSE ~ 10, minimum MSE = **7.01**



5. Results and Conclusions

Real-world Implication

The model allows consumers to predict the price of airbnb listing, based on their destinations, accommodations, desired house type. This will allow customers to better allocate their budget and make better traveling plans.

With continued improvements to the algorithm, this model benefits the hosts by giving them suggestion on pricing adjustments according to the demand during and after the pandemic.

We hope this new pricing model can help the airbnb market set price more accurately and efficiently during the pandemic, match the supply and demand, as well as supporting more hosts to survive this difficult time.

Things to Improve

- Not all the **states** have available data in our dataset, so our model is not guaranteed to be perfectly applicable in those states
- There is no information on the **years**, so we can't see the price change over years, especially before and after the breakout of COVID 19
- We only have daily price in the dataset, and no **weekly and monthly prices** are included. There might be disparities between those data because longer stays tend to have lower average price per day
- **Total floor area** of each listing is not included in the model, which may be an important feature for the pricing prediction
- There are considerable **missing values** in columns of ***number of reviews***, ***reviews per month*****, and *review scores rating***, so we either dropped those NAs or fill them by the average value based on host/property type/neighborhood. This may cause deviations from actual values and reduce the accuracy of our prediction model

Thank You

airbnb


Q&A